Construction of Automatic Confidence Intervals in Nonparametric Heteroscedastic Regression by a Moment-Oriented Bootstrap.

Volker Sommerfeld*

February 19, 1997

Abstract

We construct pointwise confidence intervals for regression functions. The method uses nonparametric kernel estimates and the "moment-oriented" bootstrap method of Bunke which is a wild bootstrap based on smoothed local estimators of higher order error moments. We show that our bootstrap consistently estimates the distribution of $\hat{m}_h(x_0) - m(x_0)$. In the present paper we focus on fully data-driven procedures and prove that the confidence intervals give asymptotically correct coverage probabilities.

1 Introduction

We consider the nonparametric regression model

$$Y_i = m(x_i) + \epsilon_i, \quad 1 \le i \le n, \tag{1.1}$$

where the errors ϵ_i are independent, but not necessarily identically distributed random variables with zero mean and finite central moments $\mu_2(x_i)$, $\mu_3(x_i)$ and $\mu_4(x_i)$. The nonrandom design points $x_1 < \cdots < x_n$ are assumed to be equally spaced on the unit interval $[0,1]$.

We aim at defining a confidence interval for the value $m(x_0)$ of the regression function m at some interior point $x_0 \in (0,1)$ which has a coverage probability near to a prescribed nominal level $1-\alpha$.

A usual construction method of such confidence intervals is as follows. As a starting point we take some kernel estimator $\hat{m}_h(x_0)$ of $m(x_0)$ with bandwidth $h = h(n)$. Under mild regularity conditions the standardized estimator

$$S_n = \sqrt{nh}\frac{\hat{m}_h(x_0) - m(x_0) - B_n(x_0)}{V_n^{1/2}(x_0)}$$

converges in distribution to a standard normal random variable. The unknown normalizing constants, the bias $B_n(x_0)$ and the variance $V_n(x_0)$ of $\hat{m}_h(x_0)$, have to be suitably estimated.

*Institut für Mathematik, Humboldt-Universität zu Berlin, PSF 1297, D-10099 Berlin, Germany. The research was carried out within the Sonderforschungsbereich 373 at Humboldt University Berlin. The paper was printed using funds made available by the Deutsche Forschungsgemeinschaft.

In the present paper we consider an alternative construction method that consists in estimating the distribution of the pivot

$$\tilde{S}_n = \sqrt{nh}\frac{\hat{m}_h(x_0) - m(x_0)}{\hat{V}_n^{1/2}(x_0)}$$

by a "wild bootstrap" method where

$$\hat{V}_n(x_0) = \sum_{i=1}^{n} w_{k,h}^{(i)}(x_0)\hat{\mu}_2(x_i)$$

is an estimate of the variance ot $\hat{m}_h(x_0)$. Here, $w_{k,h}^{(i)}(x_0)$ is a weight function and $\hat{\mu}_2(x_i)$ denotes an estimator of the error variance at the design point x_i which will be specified later. Such a construction method was used by Härdle & Marron (1991). Yet, the wild bootstrap distribution, that depends on estimators of $\mu_2(x_i)$ and $\mu_3(x_i)$ which have a strong variability. On the other hand, Bunke (1997) shows for linear regression models, that for fixed sample sizes the quality of the bootstrap approximation is strongly influenced by the estimators of the error moments. Moreover, his simulation studies show that for moderate and small sample sizes his moment oriented bootstrap procedure which is based on smooth "small bias" estimates of the error moments $\mu_2(x_i)$, $\mu_3(x_i)$ and $\mu_4(x_i)$ compares favourably with the usual variant of wild bootstrap.

There exists an extensive literature concerning confidence intervals for nonparametric regression functions (see e.g. Härdle & Bowman, 1988 and Hall, 1992 for i.i.d. errors and Härdle & Marron, 1991 for non i.i.d. errors). Yet, most of the available literature does not take into account the specific bandwidth choice that is necessary for practical applications. Exceptions are Neumann (1992) and Neumann (1995) who proved asymptotic rates for the coverage probability of confidence intervals obtained by second order normal approximations for non i.i.d. errors. He selected the bandwidth by "full-crossvalidation" and by the $\sqrt{n}$-consistent bandwidth selector of Härdle, Hall & Marron (1991).

In the present paper we prove the asymptotic validity of data-driven confidence intervals which are based on the moment-oriented bootstrap. Hereby we perform the bandwidth choice by the "full-crossvalidation" procedure proposed by Neumann (1992). The higher order performance of the coverage probability of these confidence intervals should be the subject of further research. Anticipating the following results, we remark that we need neither undersmoothing nor explicit bias correction as in the normal approximation case. This holds because of an implicit bias correction performed by the bootstrap. Thus, we can adapt the bandwidth to the data in a natural way. Yet, in order to obtain consistent bootstrap confidence intervals we have to make sure that the initial bias in the estimation of $\hat{m}_h(x_0)$ and its bootstrap counterpart cancel out. To ensure this, we need some more smoothness in the (implicit) estimation of this bias terms than in the estimation of the regression function $\hat{m}_h(x_0)$. For this reason it is not possible to get confidence intervals which shrink with the optimal asymptotic rate.

2 Estimation of the error moments

Results of Bunke (1997) for linear models indicate that the moments of the bootstrap error distribution should be good estimates of the true error moments. Following results of Gasser, Seifert & Wolf (1993) and Müller & Stadtmüller (1987a) he proposed consistent estimators of the second, third and fourth error moments which he obtained by smoothing local estimators which are unbiased in a local vicinity of the design point x_0 when the model is linear in this vicinity. In the present section we define these estimators and state their consistency properties.

In what follows we assume that the regression function $m(x)$ is $(k+2)$ - times differentiable with a continous $(k+2)$-th derivative on $[0,1]$ $(k \geq 2)$. We define the following local estimators of the error moments.

- ERROR VARIANCE:
$$\tilde{\mu}_2(x_i) := \frac{1}{2}(Y_i - Y_{i-1})^2$$
- THIRD MOMENT:
$$\tilde{\mu}_3(x_i) := \frac{1}{6}(2Y_i - Y_{i-1} - Y_{i+1})^3$$
- FOURTH MOMENT:
$$\tilde{\mu}_4(x_i) := \frac{1}{12}(2Y_i - Y_{i-1} - Y_{i+1})^4 - \frac{1}{8}(Y_{i+2} + Y_{i-1} - Y_{i+1} - Y_i)^4.$$

Note that one could also use a "second order difference" estimator of the error variance which is given by

$$\tilde{\tilde{\mu}}_2(x_i) := \frac{1}{6}(2Y_i - Y_{i-1} - Y_{i+1})^2.$$

Generalizing Lemma 2.1 of Müller & Stadtmüller (1987a) to higher order moments it is easily seen that these local estimators are asymptotically unbiased but not consistent because their variances do not vanish asymptotically. Yet, assuming some smoothness of the error moments we get consistency by smoothing these local estimators in an appropriate way.

We assume that the error moments $\mu_j(x)$, $j = 2,3,4$ as functions of the explanatory variable x are r-times $(r \geq 2)$ differentiable with a continuous second derivative on $[0,1]$. We perform the smoothing with the Gasser-Müller kernel estimator (see Gasser & Müller, 1979):

$$\hat{\mu}_j(x_0) = \hat{\mu}_j^{\lambda_j}(x_0) = \sum_{i=1}^{n} w_{r,\lambda_j}^{(i)}(x_0)\tilde{\mu}_i, \tag{2.1}$$

where

$$w_{r,\lambda_j}^{(i)}(x_0) := \frac{1}{\lambda_j}\int_{s_{i-1}}^{s_i} K_r\left(\frac{x_0 - u}{\lambda_j}\right) du,$$

$s_j = (x_j - x_{j-1})/2$ and K_r is some usual symmetric r-th order kernel with compact support $[-\tau, \tau]$ if $\lambda_j \leq x_0 \leq 1 - \lambda_j$ and some boundary kernel otherwise. Explicit

formulas for such boundary kernels are given in Gasser, Müller & Mammitzsch (1985). We further assume that K_r is Lipschitz continous of order γ_{K_r}. Obviously the differences $\tilde{\epsilon}_i^{[j]} := \tilde{\mu}_j(x_i) - \mu_j(x_i)$ and $\tilde{\epsilon}_j^{[j]} := \tilde{\mu}_j(x_l) - \mu_j(x_l)$ are independent random variables iff $|i-l| \geq j$.

The following lemma generalizes theorems 1 and 2 of Gasser & Müller (1979) for the moments of the above kernel smoother based on such (j-1) - dependent errors. We denote

$$\nu_j(x_i, x_l) := cov(\tilde{\epsilon}_i^{[j]}, \tilde{\epsilon}_l^{[j]})$$

and

$$\nu_j(x_i) := \nu_j(x_i, x_l) := Var\tilde{\epsilon}_i^{[j]}.$$

For the following lemma we assume that the error moments $\mu_j(x)$ $(j = 1, \ldots, 8)$ are continuous in the explanatory variable x. Note that this assumption and the continuity of the regression function m implies the continuity of $\nu_j(t,s)$ in s and t.

Lemma 2.1 *Under $\lambda_j \to 0$, $n\lambda_j \to \infty$ and the assumptions of this section holds*

$$E\hat{\mu}_j^{\lambda_j}(x_0) = \mu_j(x_0) + \frac{\lambda_j^r}{r!}\mu_j^{(r)}(x_0)\int_{-\tau}^{\tau} u^r K_r(u)\,du + O(n^{-1}) + o(\lambda_j^r), \quad (2.2)$$

$$Var\hat{\mu}_j^{\lambda_j}(x_0) = (2j-1)\frac{\nu_j(x_0)}{n\lambda_j}\int_{-\tau}^{\tau} K_r^2(u)\,du + O((n\lambda_j)^{-(1+\gamma_{K_r})}). \quad (2.3)$$

The proof of this lemma and of the following assertions are given in the appendix. From lemma 2.1 we obtain the weak consistency of the estimates $\hat{\mu}_j^{\lambda_j}(x_0)$, $j = 2, 3, 4$. Note that under stronger conditions on the error moments and assuming Lipschitz continuity of $m^{(k)}(x_0)$ we may deduce a corresponding strong consistency result following the lines of theorem 3.1 in Müller & Stadtmüller (1987a). Yet, such a stronger result is not necessary for the asymptotic validity of data-driven confidence intervals.

3 Convergence of bootstrap confidence intervals

As we noted in the introduction, we have to use a k-th order kernel K instead an optimal $(k+2)$-th order one in the estimation of the regression function $m(x_0)$ in order to ensure that the bias and its bootstrap counterpart cancel out. Thus, let K be a k-th order kernel. We estimate $m(x_0)$ by a Gasser-Müller kernel smoother. Thus we get the initial estimator

$$\hat{m}(x_0) = \hat{m}_h(x_0) = \sum_{i=1}^{n} w_{k,h}^{(i)}(x_0) Y_i$$

where a k-th order kernel is used in the smoothing. The moment oriented bootstrap is defined as follows.

- We denote by F_i the (unknown) distribution of the errors $\{\epsilon_i\}$. We approximate F_i by a bootstrap distribution $\hat{F}_{n,i}$ which has the first four central moments 0, $\hat{\mu}_{2,i}$, $\hat{\mu}_{3,i}$ and $\hat{\mu}_{4,i}$.

- Bootstrap observations are given by independent random variables (conditionally under the observations 1.1) $Y_i^* = \hat{m}_g(x_i) + \epsilon_i^*$ with $\epsilon_i^* \sim \hat{F}_n$, $i = 1, \ldots, n$. The bandwidth g will specified later.
- A bootstrap estimator $\hat{m}_{h_*}^*$ of $\hat{m}_h$ is obtained by a kernel smoothing of the bootstrap observations Y_i^*.

In this section we will prove the validity of the bootstrap for nonrandom bandwidths. This result will be extended to data-driven bandwidths in the next section. Note that for non random bandwidths $h = h_n$ with $h \to 0$ and $nh \to \infty$ for $n \to \infty$ the following asymptotic formulas are valid (see Gasser & Müller, 1979).

$$E_F \hat{m}_h(x_0) = m(x_0) + (-1)^k \frac{h^k}{k!} m^{(k)}(x_0) \int u^2 K(u)\, du + O(n^{-1}) + o(h^2) \quad (3.1)$$

$$Var_F \hat{m}_h(x_0) = \frac{\mu_2(x_0)}{nh} \int K^2(u)\, du + O((nh)^{-(1+\gamma_K)}). \quad (3.2)$$

The normalization used in the following lemma stems from the asymptotic behavior of $Var_F \hat{m}(x_0) = O(n^{-1}h^{-1})$.

Lemma 3.1 *We denote by $\Phi_{0,V(x_0)}(z)$ the normal distribution function with mean 0 variance $V(x_0)$. Then*

$$\left| P\left\{ \sqrt{nh}(\hat{m}_h(x_0) - m(x_0)) \le z \right\} - B(x_0) - \Phi_{0,V(x_0)}(z) \right| \to 0$$

where

$$B(x_0) = (-1)^k \frac{m^{(k)}(x_0)}{k!} \int u^k K(u)\, du$$

and

$$V(x_0) = \mu_2(x_0) \int K^2(u)\, du.$$

We denote by P_* the distribution of Y_i^* $(i = 1, \ldots, n)$ conditional under the observations $Y_1, \ldots, Y_n$. Furthermore, we denote the expectation with respect to P_* conditional on the observations $Y_1, \ldots, Y_n$ by E_*. Then we get from equation 3.1

$$\left| E_* \hat{m}_{h_*}^*(x_0) - \hat{m}_g(x_0) - (-1)^k \frac{h_*^k}{k!} \hat{m}_g^{(k)}(x_0) \int u^2 K(u)\, du \right| = O_P(n^{-1}) + o_P(h_*^k).$$

Hence, we should make sure that

$$\left| \hat{m}_g^{(k)}(x_0) - m^{(k)}(x_0) \right| = o_P(1)$$

in order to achieve the same asymptotic bias for the initial statistic in lemma 3.1 and their bootstrapped counterpart, respectively. According to Gasser & Müller (1984) the variance of $\hat{m}_g^{(k)}(x_0)$ is of order $O(n^{-1}g^{-(2k+1)})$ so that g has to tend slower to zero than $n^{-1/(2k+1)}$ to ensure the consistency of $\hat{m}_g^{(k)}(x_0)$. Thus, we use the optimal bandwidth g for the estimator $\hat{m}_g^{(k)}(x_0)$ of $m^{(k)}(x_0)$ which is of order $g \sim O(n^{-1/(2(k+2)+1)}) \gg O(n^{-1/(2k+1)})$.

In what follows let $n \to \infty$, $h, h_* \to 0$ and $nh, nh_* \to \infty$. Then we get

Lemma 3.2

$$\left| P_* \left\{ \sqrt{nh_*}(\hat{m}_h^*(x_0) - \hat{m}_g(x_0)) \le z \right\} - B(x_0) - \Phi_{0,V(x_0)}(z) \right| = o_P(1).$$

Summarizing Lemma 3.1 and Lemma 3.2 we obtain the following proposition.

Proposition 3.1 *Under the assumptions of the sections 2 and 3 it holds that*

$$\left| P_* \left\{ \sqrt{nh_*}(\hat{m}_{h_*}^*(x_0) - \hat{m}_g(x_0)) \le z \right\} - P \left\{ \sqrt{nh}(\hat{m}_h(x_0) - m(x_0)) \le z \right\} \right| = o_P(1).$$

A consistent estimator of $V(x_0)$ is (see 3.2 and lemma 6.1)

$$\hat{V}_n(x_0) = \hat{\mu}_2^*(x_0) \int K^2(u)\, du.$$

Thus, bootstrap confidence intervals with an asymptotic confidence level $1-\alpha$ are given by

$$I_\alpha := \left[\hat{m}_h(x_0) + \frac{\hat{V}_n}{\sqrt{nh}} \hat{b}_{\alpha/2}, \hat{m}_h(x_0) + \frac{\hat{V}_n}{\sqrt{nh}} \hat{b}_{1-\alpha/2} \right] \tag{3.3}$$

where $\hat{b}_\alpha$ denote the α-quantile of the bootstrap distribution for the pivotal statistic $\sqrt{nh_*}(\hat{m}_{h_*}^*(x_0) - \hat{m}_g(x_0))/\hat{V}_n(x_0)$.

4 Data-driven bandwidth choice

In order to obtain confidence intervals applicable in practice a data-dependent bandwidth choice for the kernel smoothers is necessary. Note that an optimal bandwidth choice for confidence intervals is not feasible because there is a tradeoff between the length and the coverage probability of such intervals. Hence, we apply a "full-crossvalidation" criterion proposed by Neumann (1992) which is an estimator of a mean integrated squared error (MISE). The name "full-crossvalidation" has been introduced in Bunke, Droge & Polzehl (1995) for a similar modification of cross-validation and its properties has been investigated in Droge(1994). The advantage of the full-crossvalidation criterion, especially for fixed sample sizes, is that one performs the minimization over the bandwidth interval $[0, 1/2]$ avoiding not well defined constants as in the case of the usual least-squares cross-validation criterion. We define

$$FCV(h) := \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{m}_h^{[i]}(x_i) \right)^2,$$

where

$$\hat{m}_h^{[i]}(x_i) := \sum_{j=1}^{n} w_h^{(j)}(x_0) Y_j^{[i]}$$

and

$$Y_j^{[i]} := \begin{cases} Y_j, & |i-j| \ge 1, \\ (Y_{i+1} - Y_{i-1})/2, & 2 \le i = j \le n-1, \\ Y_2, & i = j = 1, \\ Y_{n-1}, & i = j = n. \end{cases}$$

The data-driven bandwidth $\hat{h} = \hat{h}_n$ is defined by

$$\hat{h} := \arg \min_{h \in [0,1/2]} FCV(h).$$

Obviously, the estimator $\hat{m}_{\hat{h}}$ with random bandwidth does not have the structure of a sum of independent random variables. Hence, Esséen's inequality does not apply directly in order to derive asymptotic normality as in the proof of lemma 3.1. Therefore we show that the estimated regression function with data-driven bandwidth $\hat{h}$ is close to an estimator $\hat{m}_{h_0}$ with nonrandom bandwidth

$$h_0 := \arg \min_{h \in [0,1/2]} E\, FCV(h)$$

using the fact that h_0 is close to $\hat{h}$. We assume the existence of error moments of any order, additionally that the variance μ_2 is bounded away from zero and that

$$\int_0^1 \left[\int_{-\tau}^{\tau} K(z)(m(x+hz)\, dz - m(x) \right]^2 dx > 0 \quad for \quad h > 0$$

and

$$\int_0^1 \left[\int_{-\tau}^{\tau} K(z)(\mu_j(x+hz)\, dz - \mu_j(x) \right]^2 dx > 0 \quad for \quad h > 0,$$

$j = 2, 3, 4$.

The following lemma is the generalization of lemma 3.1 for data-dependent bandwidths.

Lemma 4.1 *Under the above assumptions the convergence*

$$\left| P\left\{ \sqrt{n\hat{h}}(\hat{m}_{\hat{h}}(x_0) - m(x_0)) \le z \right\} - B(x_0) - \Phi_{0,V(x_0)}(z) \right| = o(1)$$

holds.

Now, in order to obtain a version of the bootstrap lemma 3.2 for data-driven bandwidths we have to specify the choice for the initial bandwidth g, for the bandwidths λ_j used in the estimation of the error moments and for the bootstrap bandwidth h_*. As we have pointed out at the end of section 3 we have to choose h_* and g such that the bootstrap bias

$$\begin{aligned} \hat{B}_{h_*,g}(x_0) &:= E_* \hat{m}^*_{h_*}(x_0) - \hat{m}_g(x_0) \\ &= (-1)^k \frac{h_*^k}{k!} \hat{m}_g^{(k)}(x_0) \int u^k K(u)\, du + O(n^{-1}) + o(h_*^k) \end{aligned}$$

asymptotically coincides with $B(x_0)$. This requires that $\hat{m}_g^{(k)}(x_0)$ is an consistent estimator for $m^{(k)}(x_0)$. One possibility for such an consistent data-driven bandwidth choice for derivatives of the regression function was proposed by Müller & Stadtmüller (1987b) and generalized by Neumann (1995) for independent but not necessarily identically distributed errors. They observed that in the case of a 2-times differentiable function $m^{(k)}(x)$ a bandwidth for $\hat{m}_g^{(k)}(x_0)$ which minimizes the asymptotic MISE is given by

$$g_0 = g_0(k,2) = C_{k,2}(K_k) C(m(x_0), \mu_2(x_0) n^{-1/(2(k+2)+1)}(1 + o(1))$$

where K_k is a kernel of order k and the constant $C_{k,2}(K_k)$ does not depend on m and μ_2. On the other hand, the optimal bandwidth for an estimator of the regression function m at the point x_0 with smoothness of degree $k+2$ is of the form

$$g_0(0,k+2) = C_{0,k+2}(K_{k+2})C(m(x_0),\mu_2(x_0))n^{-1/(2(k+2)+1)}(1+o(1)).$$

Thus, we propose a data-driven bandwidth for $\hat{m}_g^{(k)}$ of the form

$$\hat{g} := \frac{C_{k,2}(K_k)}{C_{0,k+2}(K_{k+2})}\hat{g}_0(0,k+2)$$

where $\hat{g}_0(0,k+2)$ is a consistent estimator of $g_0(0,k+2)$ as the minimizer of the full-cross-validation criterion (see Neumann, 1992).

Following Bunke (1997), the bandwidth $\lambda_j = \lambda_j(n)$ in the smoothing of the local error moment estimators $\tilde{\mu}_j(x_i)$ is chosen by the following modified full-crossvalidation criterion that takes into account the (j-1) - dependence of the differences $\tilde{\epsilon}_i^{[j]}$. We define for $j = 2,3,4$

$$FCV_j(h) := \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\mu}_j(x_i) - \hat{\mu}_j^{h,[i]}(x_i)\right)^2$$

with

$$\hat{\mu}_j^{h,[i]}(x_i) := \sum_{l=1}^{n} w_{\lambda_j}^{(j)}(x_0)\tilde{\mu}_j^{[i]}(x_l)$$

and

$$\tilde{\mu}_j^{[i]}(x_l) := \begin{cases} \tilde{\mu}_j(x_i), & |i-l| \geq j, \\ (\tilde{\mu}_j(x_{i+j}) - \tilde{\mu}_j(x_{i-j}))/2, & |i-l| \leq j-1 \ \wedge \ l+1 \leq i = j \leq n-(l+1), \\ \tilde{\mu}_j(x_{i+j}), & |i-l| \leq j-1 \ \wedge \ i \leq j, \\ \tilde{\mu}_j(x_{n-(i+j)}), & |i-l| \leq j-1 \ \wedge \ i \geq n-j, \end{cases}$$

Hence,

$$\hat{\lambda}_j := \arg\min_{\lambda_j \in [0,1/2]} FCV_j(\lambda_j).$$

Finally, we choose the bandwidth $\hat{h}_*$ for the bootstrap estimator $\hat{m}_{\hat{h}_*}^*(x_0)$ by minimization of

$$FCV_*(h) := \frac{1}{n}\sum_{i=1}^{n}\left(Y_i^* - \hat{m}_h^{*[i]}(x_i)\right)^2$$

over $h \in [0,1/2]$. Then we obtain

Lemma 4.2

$$\left| P_*\left\{\sqrt{n\hat{h}_*}(\hat{m}_{\hat{h}_*}^*(x_0) - \hat{m}_{\hat{g}}(x_0)) \leq z\right\} - B(x_0) - \Phi_{0,V(x_0)}(z)\right| = o_p(1).$$

Now, lemma 4.1 and lemma 4.2 lead to the main theorem of the present paper.

We consider fully data-driven bootstrap confidence intervals defined by

$$\hat{I}_\alpha := \left[\hat{m}_{\hat{h}}(x_0) + \frac{\hat{V}_n}{\sqrt{n\hat{h}}}\hat{\hat{b}}_{\alpha/2}, \hat{m}_{\hat{h}}(x_0) + \frac{\hat{V}_n}{\sqrt{n\hat{h}}}\hat{\hat{b}}_{1-\alpha/2} \right]$$

where $\hat{b}_\beta$ denote the β-quantile of the bootstrap distribution for the pivot $\sqrt{n\hat{h}_*}(\hat{m}^*_{\hat{h}_*}(x_0) - \hat{m}_{\hat{g}}(x_0))/\hat{V}_n(x_0)$.

Theorem 4.1 *It holds*

$$\left| P_* \left\{ \sqrt{n\hat{h}_*}(\hat{m}^*_{\hat{h}_*}(x_0) - \hat{m}_{\hat{g}}(x_0)) \le z \right\} - P\left\{ \sqrt{n\hat{h}}(\hat{m}_{\hat{h}}(x_0) - m(x_0)) \le z \right\} \right| = o_P(1).$$

From theorem 4.1 follows the asymptotic validity of fully data-driven confidence intervals, that is

$$P\left(m(x_0) \in \hat{I}_\alpha \right) = 1 - \alpha + o(1).$$

5 Proofs

Proof of lemma 2.1

The equation 2.2 is proven in Gasser & Müller (1979). We give only a sketch of the proof of 2.3 because it follows the lines of that given in Gasser & Müller (1979) for the independent case. Without restriction of generality let $j = 2$, the proofs for $j = 3, 4$ are essentially the same. We use that $cov(\tilde{\epsilon}_i^{[2]}, \tilde{\epsilon}_l^{[2]}) = 0$ for $|i - l| \ge 2$ and derive that

$$\begin{aligned} & \left| Var\hat{\mu}_2^{\lambda_2}(x_0) - \frac{3\nu_2(x_0)}{n\lambda_2} \int_{-\tau}^{\tau} K_2^2(u)\, du \right| \\ = & \left| \frac{1}{\lambda_2^2} \sum_{i,l \in \mathcal{I}, |i-l| \le 1} \int_{s_{i-1}}^{s_i} \int_{s_{l-1}}^{s_l} K_2\left(\frac{x_0 - u}{\lambda_2}\right) K_2\left(\frac{x_0 - v}{\lambda_2}\right) du\, dv\, \nu_2(x_i, x_l) \right. \\ & \left. - \frac{\nu_2(x_0)}{n\lambda_2^2} \sum_{i \in \mathcal{I}} \int_{s_{i-1}}^{s_i} K_2^2\left(\frac{x_0 - u}{\lambda_2}\right) du \right| \end{aligned}$$

where $\mathcal{I} = \mathcal{I}(x_0)$ denotes the set of indices with non vanishing kernel weights and where $\#\mathcal{I} = O(nh)$ holds because of the compactness of the kernel. We recall that the continuity of $\nu_2(\cdot, \cdot)$ is implied by the continuity of $\mu_j(\cdot)$ for $j \le 4$. Thus, the application of the mean value theorem and the continuity of $\nu_2(s, t)$ completes the proof of the lemma.

Proof of lemma 3.1

We decompose

$$\begin{aligned} \sqrt{nh}(\hat{m}_h(x_0) - m(x_0)) &= \sqrt{nh} \left\{ \frac{1}{h} \sum_{i=1}^{n} \left[\int_{s_{i-1}}^{s_i} K\left(\frac{t-u}{h}\right) du Y_i - m(x_0) \right] \right\} \\ &= B_n(x_0) + V_n(x_0) \end{aligned}$$

into a bias term

$$B_n(x_0) = \sqrt{nh}\left\{\frac{1}{h}\sum_{i=1}^{n}\left[\int_{s_{i-1}}^{s_i} K\left(\frac{t-u}{h}\right) du(m(x_i) - m(x_0))\right]\right\}$$

and a stochastic term

$$V_n(x_0) = \sqrt{nh}\left\{\frac{1}{h}\sum_{i=1}^{n}\left[\int_{s_{i-1}}^{s_i} K\left(\frac{t-u}{h}\right) du\epsilon_i\right]\right\}.$$

Note that $B_n(x_0) = B(x_0) + o(1)$ follows from 3.1. To complete the proof of this lemma, we have to show that $V_n(x_0)$ converges in distribution to $N(0, V(x_0))$. With the notation

$$W_{hi}(x_0) := \sqrt{\frac{n}{h}}\int_{s_{i-1}}^{s_i} K\left(\frac{x_0 - u}{h}\right) du.$$

we obtain

$$V_n(x_0) = \sum_{i=1}^{n} W_{hi}(x_0)\epsilon_i$$

and therefore

$$EV_n^2(x_0) = VarV_n(x_0) = \sum_{i=1}^{n} W_{hi}^2(x_0)\mu_2(x_i)$$

and

$$EV_n^3(x_0) = \sum_{i=1}^{n} W_{hi}^3(x_0)\mu_3(x_i).$$

Hence it follows that

$$\frac{EV_n^3(x_0)}{(EV_n^2(x_0))^{3/2}} = O(n^{-1/2}). \tag{5.1}$$

Now, applying Esséen's inequality for independent but not necessarily identically distributed random variables (see e.g. Petrov, 1975, S. 111) we get from 5.1

$$V_n(x_0) \xrightarrow{d} N(0, V(x_0))$$

which completes the proof of the lemma.

Proof of lemma 3.2

Analogeously to the proof of lemma 3.1 we decompose

$$\sqrt{nh}(\hat{m}_h^*(x_0) - \hat{m}_g(x_0)) = \tilde{B}_n(x_0) + \tilde{V}_n(x_0)$$

with

$$\tilde{B}_n(x_0) = \sqrt{nh}\left\{\frac{1}{h}\sum_{i=1}^{n}\left[\int_{s_{i-1}}^{s_i} K\left(\frac{t-u}{h}\right) du(\hat{m}_g(t_i) - \hat{m}_g(x_0))\right]\right\}$$

and

$$\tilde{V}_n(x_0) = \sqrt{nh}\left\{\frac{1}{h}\sum_{i=1}^{n}\left[\int_{s_{i-1}}^{s_i} K\left(\frac{t-u}{h}\right) du\epsilon_i^*\right]\right\}.$$

Now, according to 3.1, we have

$$\tilde{B}_n(x_0) \asymp \frac{\hat{m}_g^{(k)}(x_0)}{k!} \int u^k K_k(u)\, du.$$

Thus, because of $\hat{m}_g^{(k)}(x_0) = m^{(k)}(x_0) + o_P(1)$ it follows that $\tilde{B}_n(x_0) = B(x_0) + o_P(1)$. Now we deduce from 3.2 and the second part of the proof of lemma 3.1 that

$$\left| P_* \left\{ \tilde{V}_n(x_0) \le z \right\} - \Phi_{0, \lim_{n\to\infty} \hat{\mu}_2(x_0) \int K^2(u)\, du}(z) \right| = o_P(1).$$

This, together with $\hat{\mu}_2(x_0) = \mu_2(x_0) + o_P(1)$ gives

$$\left| P_* \left\{ \tilde{V}_n(x_0) \le z \right\} - \Phi_{0, V(x_0)}(z) \right| = o_P(1).$$

which completes the proof of lemma 3.2.

Proof of lemma 4.1

Note that

$$\hat{m}_{\hat{h}}(x_0) = \sum_{i=1}^n w_{\hat{h}}^{(i)}(x_0) m(x_i) + \sum_{i=1}^n w_{\hat{h}}^{(i)}(x_0) \epsilon_i.$$

At first, we consider the stochastic part $\sum_{i=1}^n w_{\hat{h}}^{(i)}(x_0)\epsilon_i$. Let h_0 be the minimizer of $EFCV(h)$. We define a grid of lattice points $H_n' := \{h_1, \ldots, h_{m_n}\} \subset H_n$. Then, applying Markov's and Whittle's inequalities, for each $h_j \in H_n'$

$$\begin{aligned}
& P\left(\left| \sum_{i=1}^n [w_{k,h_j}^{(i)}(x_0) - w_{k,h_0}^{(i)}(x_0)]\epsilon_i \right| > n^{\delta} \| w_{k,h_j}(x_0) - w_{k,h_0}(x_0) \|_{L_2} \right) \\
\le & \frac{E \left| \sum_{i=1}^n [w_{k,h_j}^{(i)}(x_0) - w_{k,h_0}^{(i)}(x_0)]\epsilon_i \right|^k}{n^{\delta k} \| w_{k,h_j}(x_0) \|_{L_2}^k} \\
\le & C_k n^{-\delta k} \\
= & O(n^{-\lambda}) \qquad (5.2)
\end{aligned}$$

holds when $k \ge \frac{\lambda}{\delta}$. Here we take k sufficiently large to obtain for any small $\delta > 0$ a good rate of convergence $n^{-\lambda}$. On the other hand, the Bonferroni inequality and 5.2 give

$$\begin{aligned}
& P\left(\exists h_j \in H_n' : \left| \sum_{i=1}^n [w_{k,h_j}^{(i)}(x_0) - w_{k,h_0}^{(i)}(x_0)]\epsilon_i \right| > n^{\delta} \| w_{k,h_j}(x_0) - w_{k,h_0}(x_0) \|_{L_2} \right) \\
\le & \sum_{j=1}^{m_n} P\left(\left| \sum_{i=1}^n [w_{k,h_j}^{(i)}(x_0) - w_{k,h_0}^{(i)}(x_0)]\epsilon_i \right| > n^{\delta} \| w_{k,h_j}(x_0) - w_{k,h_0}(x_0) \|_{L_2} \right) \\
= & m_n O(n^{-\lambda}) = O(n^{-\alpha}) \qquad (5.3)
\end{aligned}$$

where $\# H_n' = n^{\beta}$ is the cardinality of the grid.

Now, we denote by $h(\hat{h}) \in H_n'$ the point of the grid which is the closest to $\hat{h}$. Then, obviously

$$|\hat{h} - h(\hat{h})| \le n^{-\beta} \qquad (5.4)$$

holds in probability because of $H_n' \subset [0, 1/2]$. On the other hand, we derive by Bonferroni's and Markov's inequality that

$$\begin{aligned} P(\exists i \in \{1, \ldots, n\} : |\epsilon_i| > n^\delta) &\leq \sum_{i=1}^{n} P(|\epsilon_i| > n^\delta) \\ &\leq \sum_{i=1}^{n} \frac{E|\epsilon_i|^k}{n^{\delta k}} \\ &= n^{1-\delta k} E|\epsilon_i|^k = O(n^{-\alpha}) \end{aligned} \tag{5.5}$$

for sufficiently large k. From 5.5 follows

$$\begin{aligned} & P\left(\left|\sum_{i=1}^{n}[w_{k,\hat{h}}^{(i)}(x_0) - w_{k,h(\hat{h})}^{(i)}(x_0)]\epsilon_i\right| > n^\delta \sum_{i=1}^{n} |w_{k,\hat{h}}^{(i)}(x_0) - w_{k,h(\hat{h})}^{(i)}(x_0)|\right) \\ \leq\ & P\left(\max_{i=1,\ldots,n} |\epsilon_i| \sum_{i=1}^{n} |w_{k,\hat{h}}^{(i)}(x_0) - w_{k,h(\hat{h})}^{(i)}(x_0)| > n^\delta \sum_{i=1}^{n} |w_{k,\hat{h}}^{(i)}(x_0) - w_{k,h(\hat{h})}^{(i)}(x_0)|\right) \\ =\ & P\left(\max_{i=1,\ldots,n} |\epsilon_i| > n^\delta\right) \\ =\ & O(n^{-\alpha}). \end{aligned} \tag{5.6}$$

Hence with probability $\geq 1 - O(n^{-\alpha})$ we have

$$\begin{aligned} \left|\sum_{i=1}^{n}[w_{k,\hat{h}}^{(i)}(x_0) - w_{k,h(\hat{h})}^{(i)}(x_0)]\epsilon_i\right| &\leq n^\delta \sum_{i=1}^{n} |w_{k,\hat{h}}^{(i)}(x_0) - w_{k,h(\hat{h})}^{(i)}(x_0)| \\ &= n^\delta \sum_{i=1}^{n} O\left(\frac{\hat{h} - h(\hat{h})}{h(\hat{h})} \frac{1}{nh(\hat{h})}\right) \\ &= O\left(n^\delta \frac{1}{h(\hat{h})^2} |\hat{h} - h(\hat{h})|\right) \\ &= O\left(n^\delta \frac{1}{h(\hat{h})^2} n^{-\beta}\right) \\ &= O(n^{-\kappa}), \end{aligned} \tag{5.7}$$

with $\kappa \geq 1$. Hereby the first equality follows from lemma 6.1, the last from inequality 5.4 assuming $\beta \geq 3$.

Furthermore, we deduce from 5.3, 5.7 and lemma 6.1 that with probability $\geq 1 - O(n^{-\alpha})$

$$\begin{aligned} & \left|\sum_{i=1}^{n}[w_{k,\hat{h}}^{(i)}(x_0) - w_{k,h_0}^{(i)}(x_0)]\epsilon_i\right| \\ \leq\ & \left|\sum_{i=1}^{n}[w_{k,\hat{h}}^{(i)}(x_0) - w_{k,h(\hat{h})}^{(i)}(x_0)]\epsilon_i\right| + \left|\sum_{i=1}^{n}[w_{k,h(\hat{h})}{}^{(i)}(x_0) - w_{k,h_0}^{(i)}(x_0)]\epsilon_i\right| \end{aligned}$$

$$
\begin{aligned}
&\leq O(n^{-1}) + n^{\delta}\|w_{k,h(\hat{h})} - w_{k,h_0}\| \\
&= O\left(n^{-1} + \frac{h(\hat{h}) - h_0}{h_0}(nh_0)^{-1/2}n^{\delta}\right). \qquad (5.8)
\end{aligned}
$$

Note that Neumann (1992), lemma 4.1 proved the relation

$$\frac{h(\hat{h}) - h_0}{h_0} = O_P(n^{-1/10}).$$

Thus, for $\delta < 1/10$ it follows from 5.8

$$\sqrt{nh_0}\left(\sum_{i=1}^{n}\left[w^{(i)}_{k,\hat{h}}(x_0) - w^{(i)}_{k,h_0}(x_0)\right]\epsilon_i\right) \xrightarrow{P} 0. \qquad (5.9)$$

For the bias term we derive by a Taylor expansion of $w^{(i)}_h(x_0)$ as a function of h

$$
\begin{aligned}
&\left|\sum_{i=1}^{n}\left[w^{(i)}_{k,\hat{h}}(x_0) - w^{(i)}_{k,h_0}(x_0)\right]m(x_i)\right| \\
= &\left|(\hat{h} - h_0)\sum_{i=1}^{n}\frac{d}{dh}w^{(i)}_{k,h}(x_0)\Big|_{h=\tilde{h}}m(t_i)\right|, \qquad (5.10)
\end{aligned}
$$

with $\tilde{h}$ between $\hat{h}$ and h_0. Hence, from

$$\sum_{i=1}^{n}\frac{d}{dh}w^{(i)}_{k,h}(x_0) = \frac{d}{dh}Bias(\hat{m}_h(x_0)) = O(h) \qquad (5.11)$$

follows

$$\sum_{i=1}^{n}\left[w^{(i)}_{k,\hat{h}}(x_0) - w^{(i)}_{k,h_0}(x_0)\right]m(t_i) = O\left(\frac{h(\hat{h}) - h_0}{h_0}h_0^2\right) = o_p(h^2). \qquad (5.12)$$

Summing up 5.9 and 5.12 we get

$$\sqrt{nh_0}\,(\hat{m}_{\hat{h}}(x_0) - \hat{m}_{h_0}(x_0)) \xrightarrow{P} 0. \qquad (5.13)$$

Finally, the application of lemma 3.1, equation 5.13 and Slutsky's theorem makes the proof of this lemma complete.

Proof of lemma 4.2

The proof of Lemma 4.2 is similar in spirit to that of lemma 4.1. The only real difference is that we have to show

$$|\hat{\mu}_j^{\hat{\lambda}_j}(t) - \mu_j(t)| \xrightarrow{P} 0 \qquad (5.14)$$

where

$$\hat{\lambda}_j := \arg\min_{\lambda_j \in [0,1/2]} FCV_j(\lambda_j).$$

Yet, by applying lemma 6.3 of this paper, lemma 4.1 in Neumann (1992) is easily generalized for (j-1) - dependent random error terms, that is

$$\frac{\hat{\lambda_j} - \lambda_{j,0}}{\lambda_{j,0}} = O_P(n^{-1/10})$$

for λ_j chosen as above and

$$\lambda_{j,0} := \arg \min_{\lambda_j \in [0,1/2]} E\, FCV_j(\lambda_j).$$

Furthermore, we conclude from 2.3 that for nonrandom bandwidth λ_j

$$|\hat{\mu}_j^{\lambda_j}(t) - \mu_j(t)| \xrightarrow{P} 0. \tag{5.15}$$

Thus, we easily complete the proof of this lemma arguing along the lines of the proof of lemma 4.1.

6 Some technical lemmas

Lemma 6.1 *Assume for some bandwidths g and h that $\frac{g-h}{h} \to 0$. Then*

$$|w_g^{(i)}(x_0) - w_h^{(i)}(x_0)| = O\left(\frac{g-h}{h}\frac{1}{nh}\right) \tag{6.1}$$

Proof.

Recall that

$$w_g^{(i)}(x_0) = \frac{1}{g}\int_{s_{i-1}}^{s_i} K\left(\frac{t-u}{g}\right) du.$$

Taylor expansions of g^{-1} at h and of $K((x_0-u)/g)$ give

$$\begin{aligned} w_g^{(i)}(x_0) &= w_h^{(i)}(x_0) \\ &\quad + \int_{s_{i-1}}^{s_i} \left(\frac{g-h}{\tilde{g}^2}\left(K'\left(\frac{t-u}{\check{g}}\right)\left(\frac{g-h}{\tilde{g}^2} - \frac{1}{h}\right) - K\left(\frac{t-u}{h}\right)\right)\right) du \end{aligned}$$

where $\check{g}$ and $\tilde{g}$ are between g and h. Thus, by the mean value theorem with $s_{i-1} \le \eta \le s_i$ we obtain

$$\begin{aligned} & w_g^{(i)}(x_0) - w_h^{(i)}(x_0) \\ = \;& (s_i - s_{i-1})\frac{g-h}{\tilde{g}^2}\left(K'\left(\frac{t-\eta}{\check{g}}\right)\left(\frac{g-h}{\tilde{g}^2} - \frac{1}{h}\right) - K\left(\frac{t-\eta}{h}\right)\right) \\ = \;& O\left(\frac{g-h}{h}\frac{1}{nh}\right). \end{aligned}$$

The following 2 lemmas give a generalization of Whittles inequalities to (m-1) - dependent random variables.

Lemma 6.2 (Generalization of Theorem 1 in Whittle, 1960) *Assume that the random variables $Z_1, \ldots, Z_n$ are (m-1) - dependent, that is Z_i and Z_{i+m} are independent. We assume further that $P(Z_j = -1) = P(Z_j = 1) = 1/2$. Then, for $s \geq 2$*

$$E \left| \sum_{j=1}^{n} b_j Z_j \right|^s \leq C_m(s) \left(\sum_{j=1}^{n} b_j^2 \right)^{s/2}$$

holds.

Proof.

Without loss of generality we restrict ourselves to $m = 2$. The only difference to the proof in the case of independent random variables is that there is another constant $C_m(s)$ instead of $C(s)$ in Whittles theorem.

Let us define the sets of indicees $\mathcal{J}_1$ and $\mathcal{J}_2$ such that the random variables that are indexed by one index set are independent. Then

$$\begin{aligned} \left(\sum_{j=1}^{n} Z_j \right)^s &= \left(\sum_{j \in \mathcal{J}_1} Z_j + \sum_{j \in \mathcal{J}_2} Z_j \right)^s \\ &= \sum_{k=0}^{s} \binom{s}{k} \left(\sum_{j \in \mathcal{J}_1} Z_j \right)^k \left(\sum_{j \in \mathcal{J}_2} Z_j \right)^{s-k}, \end{aligned}$$

hence

$$\begin{aligned} E \left| \sum_{j=1}^{n} Z_j \right|^s &\leq \sum_{k=0}^{s} \binom{s}{k} E \left| \sum_{j \in \mathcal{J}_1} Z_j \right|^k \left| \sum_{j \in \mathcal{J}_2} Z_j \right|^{s-k} \\ &\leq \sum_{k=0}^{s} \binom{s}{k} C(k) C(s-k) =: C_2(k). \end{aligned}$$

Thereby the bounds by the constants $C(\cdot)$ are valid according to theorem 1 of Whittle (1960) so that the lemma is proved.

Now, following the lines of the proof of theorem 2 in Whittle (1960) with this new constant we easily deduce the followig lemma for (m-1) - dependent random variables $X_1, \ldots, X_n$ with expectation 0.

Lemma 6.3 (Generalization of Theorem 2 in Whittle, 1960) *Let $X := (X_1, \ldots, X_n)'$, $s \geq 2$ and $\gamma_j(s) := (E|X_j|^s)^{1/s}$. Then for any matrices $A = (a_{jk})$ and vectors $B = (b_j)$*

$$\begin{aligned} E|X'B|^s &\leq 2^s C_m(s) \left(\sum_{j=1}^{n} b_j^2 \gamma_j^2(s) \right)^{s/2}, \\ E|X'AX - EX'AX|^s &\leq 2^{3s} C_m(s) (C_m(2s))^{1/2} \left(\sum_{j=1}^{n} \sum_{k=1}^{n} a_{jk}^2 \gamma_j^2(2s) \gamma_k^2(2s) \right)^{s/2} \end{aligned}$$

if we assume the existence of the right-hand sides of the inequalities.

References

Bunke, O. (1997). Bootstrapping in Heteroscedastic Regression Situations. *Discussion Paper*, Sonderforschungsbereich 373, Humboldt University, Berlin, to appear.

Bunke, O., Droge, B. & Polzehl, J. (1995). Model Selection, Transformations and Variance Estimation in Nonlinear Regression. *Discussion Paper* **52**, Sonderforschungsbereich 373, Humboldt-Universität, Berlin.

Droge, B. (1994). Some Comments on Cross-Validation. *Discussion Paper* **7**, Sonderforschungsbereich 373, Humboldt University, Berlin.

Gasser, T. & Müller, H.G. (1979). Kernel Estimation of Regression Functions, in Smoothing Techniques in Curve Estimation. *Lecture Notes in Mathematics* **757**, 23-68.

Gasser, T. & Müller, H.G. (1984). Estimating Regression Functions and Their Derivatives by the Kernel Method. *Scand. J. Statist.* **11**, 171-185.

Gasser, T. & Müller, H.G. & Mammitzsch, V. (1985). Kernels for Nonparametric Curve Estimation. *J. Roy. Statist. Soc. Ser. B* **47**, 238-252.

Gasser, T. , Seifert, B. & Wolf, A. (1993). Nonparametric Estimation of Residual Variance Revisited. *Biometrika* **80**, 375-383.

Härdle, W. & Bowman, A. (1988). Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands. *J. Amer. Statist. Assoc.* **83**, 102-110.

Härdle, W., Hall, P. & Marron, J. S. (1991). Regression Smoothing Parameters that are not far from their Optimum. *J. Amer. Statist. Assoc.* **87**, 227-233.

Härdle, W. & Marron, J. S. (1991). Bootstrap Simultaneous Error Bars in Nonparametric Regression. *Ann. Statist.* **19**, 778-796.

Hall, P. (1992). On Bootstrap Confidence Intervals in Nonparametric Regression. *Ann. Statist.* **20**, 695-711.

Müller, H.-G. & Stadtmüller, W. (1987a). Estimation of Heteroscedasticity in Regression Analysis. *Ann. Statist.* **15**, 610-625

Müller, H.-G. & Stadtmüller, W. (1987b). Variable Bandwidth Kernel Estimators of Regression Curves. *Ann. Statist.* **15**, 182-201.

Neumann, M. H. (1992). On Completely Data-driven Pointwise Confidence Intervals in Nonparametric Regression. *Rapport Technique* **92-02**, INRA, Dépt. de Biométrie, Jouy-en-Josas, France.

Neumann, M. H. (1995). Automatic Bandwidth Choice and Confidence Intervals in Nonparametric Regression. *Ann. Statist.* **23**, 1937-1959.

Petrov, V. (1975). *Sums of Independent Random Variables.* Springer, New York.

Whittle, P. (1960). Bounds for the Moments of Linear and Quadratic Forms in Independent Variables. *Theory Probab. Appl.* **5**, 302-305.